QIIME for Microbiome Analysis: Amplicon Design and Sequencing

ROY J. CARVER BIOTECHNOLOGY CENTER
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Chris Wright
Assistant Director of DNA Services

Alvaro Hernandez
Director of DNA Services
Outline:

- Overview of 16S sequencing and applications
- Construction of 16S and other targeted amplicons
- Sequencing Results
- Pre-QIIME questions and concepts
Microbial Community Sequencing—3 main questions:

1) Who is there?
   → 16s (or other) amplicon sequencing
   → metagenomic DNA sequencing

2) What can they do?
   → metagenomic DNA sequencing

3) What are they doing now?
   → metatranscriptome (RNA) sequencing
Applications for Microbiome Sequencing:

- Earth Microbiome
- Human Microbiome
- Disease states (obesity, diabetes)
- Home Microbiome
- Environmental surveillance, damage assessment and habitat recovery
  - Chemical waste cleanup
  - Oil Spills
- Evolutionary Studies
- Soil/plant function impacts
- Deep well oil drilling & recovery
- Coral reef disease
- Exercise
- Companion/farm animal diets
- Waste and wastewater treatment
- Farming
- Infections Disease surveillance
Targeted Amplicon Sequencing:

Requires construction of an Illumina-compatible library:

- **Design and/or identify primers**
  - 16S rRNA (bacterial/archaeal)
  - 18S (eukaryotic)
  - ITS (fungi)
  - Any other gene (amoA, rpoB, nosZ, nrfA, etc.)

For each sample:

- **PCR**
- **Pool**
- **Submit for sequencing**
- **Cleanup/gel purify**
- **Quantitate/Qualitate**
Fluidigm Multi-Targeted Amplicon Library Preparation:

- Submit gDNA to the Biotechnology Center

- Each plate amplifies 48 samples against up to 48 primers
  - 10x multiplex capable ➔ over 23,000 PCR reactions

- For microbial communities (16S, ITS, 18S, Archaea, protozoan, etc):
  - Typical project is 2-6 primer sets.
  - Up to 24 primer sets per plate.

- 384 barcodes available
  - Pool 8 plates per sequencing run
Fluidigm PCR Construct:

Region of Interest:

Final Construct:

Read 1

Index Read

Read 2
Project Design: Pre-Sequencing & Sequencing Questions

- What primer-sets to use??
- How many samples??
- How much sequencing do I need??

- DNA Services/Functional Genomics lab can offer:
  - Primer set choices and testing.
  - Options for sequencing.
  - Costs across different designs.

- Researcher should examine:
  - What primers are researchers in this field using?
  - Are you comparing to previously published data?
  - How much sequencing will your funding allow?
### Primer Options:
- Ever-growing list!
- Includes 16S, 18S, ITS, eukaryotic, gene-specific targets, more.
- Functional Genomics lab will order and test any primer-set that is not on the list.

<table>
<thead>
<tr>
<th>Set</th>
<th>Primer Set Name</th>
<th>Forward Primer Name</th>
<th>Forward Primer Sequence</th>
<th>Reverse Primer Name</th>
<th>Reverse Primer Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>V3_V5_F357_R926</td>
<td>V3_F357</td>
<td>CCTACGGGAGGGACACGAG</td>
<td>V3_V5_R926</td>
<td>CGTCATTACMTTTRAGT</td>
</tr>
<tr>
<td>2</td>
<td>V1_V3_F28_R519</td>
<td>V1_F28</td>
<td>GAGTTTGATCNTGGCTG</td>
<td>V1_V3_R519</td>
<td>GTTTACNGCGGCKGCTG</td>
</tr>
<tr>
<td>3</td>
<td>Arch349F_806R</td>
<td>Arch349F</td>
<td>GYGCASCAGKCGMGAAW</td>
<td>Arch806R</td>
<td>GGACTACVSGGGTATACTAAT</td>
</tr>
<tr>
<td>4</td>
<td>F566Euk_R1200</td>
<td>F566Euk</td>
<td>CAGCAGCCGCGGTAATTCC</td>
<td>R1200Euk</td>
<td>CCCGTGTGAGTCAAATTAAGC</td>
</tr>
<tr>
<td>5</td>
<td>ITS1_ITS4</td>
<td>ITS1</td>
<td>TCCGTAGGTGAAACCTGG</td>
<td>ITS4</td>
<td>TCCTCCGCTTTATTGATATGC</td>
</tr>
<tr>
<td>6</td>
<td>ITS3_ITS4</td>
<td>ITS3</td>
<td>GCATCGATGAAGAAGCGCAG</td>
<td>ITS4</td>
<td>TCCTCCGCTTTATTGATATGC</td>
</tr>
<tr>
<td>7</td>
<td>Euk_1391f_EukBr</td>
<td>Euk_1391f</td>
<td>GTACACACCGCCGCAGC</td>
<td>EukBr</td>
<td>TGATCTTCTGCAGGTTTCACCTAC</td>
</tr>
</tbody>
</table>
Sequencing Options:

- **MiSeq v3**
  - Options for 250nt or 300nt read length
  - Output of 1M, 10M, or 20M paired-reads per run

- **HiSeq 2500**
  - Options for 150nt or 250nt read length
  - Output of 100-150M paired-reads per lane

- Cost from $1,200 to $4,050 per run
Outline:

- Overview of 16S sequencing and applications
- Construction of 16S and other targeted amplicons
- Sequencing Results
- Pre-QIIME questions and concepts
**Sequencing Results:**

- Data (fastq) posted to password-protected FTP site.
- Six-page Excel report.
- Quality Score graphs for all samples & primer sets.
- Run data (fastq) processed in 4 ways:
  - Unsorted
  - Demultiplexed
  - Primer Sorted
  - Primer Sorted & Demultiplexed
Sequencing Results, Run Metrics tab:

1) **Unsorted data**: This is bulk read data. It is not sorted, trimmed, demultiplexed, or processed in any way other than to remove PhiX reads.

   ➔ *Shows total number of reads generated for the project.*

2) **Demultiplexed data**: This data is sorted by index and PhiX reads are removed. No other processing is done on this data set.

   ➔ *Shows how well individual samples worked across all the primer-sets.*
Sequencing Results, Primer Sorted Metrics Tab:

3) **Primer Sorted data**: This data is sorted by PCR-specific primers and PhiX reads are removed. No other processing is done on this data set.

   ➔ *Shows number of reads for each primer-set*

4) **Primer Sorted and Demultiplexed data**: This data is sorted by PCR-specific primers and then by index. PhiX reads are removed. No other processing is done on this data set.

   ➔ *This is the fully sorted and demultiplexed data for each sample and each primer.*
Sequencing Results, Barcode File tab:

- Shows the sample ID and associated barcode (index).
- You need this for your mapping file in QIIME.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Barcode Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CACGAAGAGC</td>
</tr>
<tr>
<td>7</td>
<td>CATACCTGAT</td>
</tr>
<tr>
<td>13</td>
<td>GACGTGCTTC</td>
</tr>
<tr>
<td>19</td>
<td>ATGTGGAGT</td>
</tr>
<tr>
<td>25</td>
<td>TCTGGTCTCA</td>
</tr>
<tr>
<td>31</td>
<td>AGGTAAAGGG</td>
</tr>
<tr>
<td>37</td>
<td>TCCTGACAGA</td>
</tr>
<tr>
<td>43</td>
<td>GCACGTGTC</td>
</tr>
<tr>
<td>2</td>
<td>ACCATGAGTC</td>
</tr>
<tr>
<td>8</td>
<td>AATGCAGTGT</td>
</tr>
<tr>
<td>14</td>
<td>TATGGTGGGA</td>
</tr>
<tr>
<td>20</td>
<td>ACTCAGTTAC</td>
</tr>
<tr>
<td>26</td>
<td>AAGTGCAGATG</td>
</tr>
<tr>
<td>32</td>
<td>CCACAGAGTG</td>
</tr>
<tr>
<td>38</td>
<td>AGTGGTGATC</td>
</tr>
<tr>
<td>44</td>
<td>ACTTCTTAGC</td>
</tr>
<tr>
<td>3</td>
<td>GCCACATATA</td>
</tr>
</tbody>
</table>
Sequencing Results, Sample-Primer tab:

- Shows trace files of final Fluidigm Pools
- Shows graphic of Fluidigm amplicon structure.
Sequencing Results, Sample-Primer tab:

- Lists locus-specific primers used on the genomic DNA.
- Lists CS1 and CS2 Fluidigm-specific pad sequences.
- Lists the Ilumina i5 and i7 end sequences.

<table>
<thead>
<tr>
<th>Primer target:</th>
<th>Primer name:</th>
<th>Locus-specific primer sequence:</th>
<th>Expected length(nt) of amplified region including locus-specific primer sequences:</th>
<th>Expected total fragment length(nt) with index and Illumina adaptors:</th>
</tr>
</thead>
<tbody>
<tr>
<td>16S_V3-V5</td>
<td>16s v3-v5-F357</td>
<td>5’-CCTACGGGAGGGAGCAGCAG</td>
<td>570</td>
<td>676</td>
</tr>
<tr>
<td></td>
<td>16s v3-v5-R926</td>
<td>5’-CCGTCATTCMTTTRAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITS1-ITS4</td>
<td>ITS1F-10</td>
<td>5’TCCGTTAGGTAACCTGCGG</td>
<td>474+ variable</td>
<td>580+ variable</td>
</tr>
<tr>
<td></td>
<td>ITS4R-10</td>
<td>5’TCTCTCGGTTATGATATGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMV4.5NF_AMDGR</td>
<td>AMV4.5NF_AMDGR</td>
<td>5’-AAGCTCGTAGTTGAATTTGCG</td>
<td>~254 (extrapolated from final product on bioanalyzer)</td>
<td>~360 (from bioanalyzer)</td>
</tr>
<tr>
<td></td>
<td>AMDGR</td>
<td>5’-CCCAGATCCCTATGATATGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2_R6</td>
<td>F2</td>
<td>5’-TGYGAYCCIAAGCCIGA</td>
<td>~384 (extrapolated from final product on bioanalyzer)</td>
<td>~490 (from bioanalyzer)</td>
</tr>
<tr>
<td></td>
<td>R6</td>
<td>5’-TCIGGIGARATGATG</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sequencing Results, FTP-information tab:

- Shows how to log into the FTP site and download data.
- Do not download to your PC or laptop—files are generally too big.
- Do download to a server.
- Do make at least 2 copies of your data (give one to your PI!)
Outline:

- Overview of 16S sequencing and applications
- Construction of 16S and other targeted amplicons
- Sequencing Results
- Pre-QIIME questions and concepts
Data Handling: Processing from the instrument

• QIIME favors bulk (non-demultiplexed) files
• You will use the Primer Sorted fastq files.
• Each primer set must be processed separately.
Data Handling: “Other” Sequencing files:

• To make demultiplexed files compatible for QIIME:
  • `add_qiime_labels.py` – Takes a directory, a metadata mapping file, and a column name that contains the fasta file names that SampleIDs are associated with, combines all files that have valid fasta extensions into a single fasta file, with valid QIIME fasta labels.

• Dual-indexed samples:
  • These CAN be processed in QIIME.
  • Index read sequences are merged to make one fastq file.
Data Handling: Trimming Data

- Posted fastqs are **not** trimmed for adaptors or primers.
- Posted fastqs are **not** trimmed for quality.
- You should trim adaptors and primers from reads before doing any data processing.
- **If** sequencing read includes forward locus-specific primer, trim it after data is binned by primer.
- **If** sequencing read includes reverse locus-specific primer, trim it **and** all sequence after it.
Data Handling: Merging Overlapping Paired-Reads

- Can only pair if reads overlap—check read lengths against the length of your PCR product.
- Pair after adaptor & primer & quality trimming!
- Multiple options within and outside of QIIME (Pear, join_paired_ends.py)
Data Handling: Starting to QIIME:

- **Mapping File**
  - contains all metadata, sample names, indexes, etc.
  - [http://qiime.org/documentation/file_formats.html](http://qiime.org/documentation/file_formats.html)
  - Minimum columns: SampleID, BarcodeSequence, LinkerPrimerSequence, and Description.
  - Leave column empty if no data, but keep header!

<table>
<thead>
<tr>
<th>SampleID</th>
<th>BarcodeSequence</th>
<th>LinkerPrimerSequence</th>
<th>Treatment</th>
<th>DOB</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC.354</td>
<td>AGCACGAGCCTA</td>
<td>YATGCTGCCCTCCGAGGAGT</td>
<td>Control</td>
<td>20061218</td>
<td>Control_mouse__I.D._354</td>
</tr>
<tr>
<td>PC.355</td>
<td>AACCTCGTCGATG</td>
<td>YATGCTGCCCTCCGAGGAGT</td>
<td>Control</td>
<td>20061218</td>
<td>Control_mouse__I.D._355</td>
</tr>
<tr>
<td>PC.356</td>
<td>ACAGACCACTCA</td>
<td>YATGCTGCCCTCCGAGGAGT</td>
<td>Control</td>
<td>20061126</td>
<td>Control_mouse__I.D._356</td>
</tr>
<tr>
<td>PC.481</td>
<td>ACCAGCAGTACTA</td>
<td>YATGCTGCCCTCCGAGGAGT</td>
<td>Control</td>
<td>20070314</td>
<td>Control_mouse__I.D._481</td>
</tr>
<tr>
<td>PC.593</td>
<td>ACAGAGACTTGT</td>
<td>YATGCTGCCCTCCGAGGAGT</td>
<td>Control</td>
<td>20071210</td>
<td>Control_mouse__I.D._593</td>
</tr>
<tr>
<td>PC.607</td>
<td>AACTGTCCTGATC</td>
<td>YATGCTGCCCTCCGAGGAGT</td>
<td>Fast</td>
<td>20071112</td>
<td>Fasting_mouse__I.D._607</td>
</tr>
<tr>
<td>PC.634</td>
<td>ACAGATCGCGCT</td>
<td>YATGCTGCCCTCCGAGGAGT</td>
<td>Fast</td>
<td>20080116</td>
<td>Fasting_mouse__I.D._634</td>
</tr>
<tr>
<td>PC.635</td>
<td>ACCGCAGAGCTA</td>
<td>YATGCTGCCCTCCGAGGAGT</td>
<td>Fast</td>
<td>20080116</td>
<td>Fasting_mouse__I.D._635</td>
</tr>
<tr>
<td>PC.636</td>
<td>ACCTGAGTGTCA</td>
<td>YATGCTGCCCTCCGAGGAGT</td>
<td>Fast</td>
<td>20080116</td>
<td>Fasting_mouse__I.D._636</td>
</tr>
</tbody>
</table>

- **Fastq files**, already primer/quality trimmed and reads merged.
Getting Started with Sequencing:

1) Chemistries, library kits, yields, protocols, turnaround times, instruments, etc. change every few weeks, months, years...

   Read papers, but get in touch with us **before** you start. **Things change....FAST!**

2) **Submission form** for genomic DNA samples for Fluidigm amplification and sequencing are available through the Functional Genomics Lab

   ➔ call or email for an appointment with us.
THANKS!

Functional Genomics Lab (Fluidigm)
Mark Band, Director
markband@illinois.edu
356 ERML
217-244-3930

www.biotech.illinois.edu/
functionalgenomics

DNA Services Lab
(Library Prep & Sequencing)
Alvaro Hernandez, Director
aghernan@illinois.edu
329 ERML
217-244-3480

Chris Wright, Assistant Director
clwright@illinois.edu
334 ERML
217-333-4372

www.biotech.illinois.edu/htdna